Preprint typeset HYPER VERSION



# A method for statistical comparison of histograms

**S. Bityukov[a]*, N. Krasnikov[b], A. Nikitenko[c] and V. Smirnova[a]**

[a]*Institute for high energy physics,*
*142281 Protvino, Russia*
[b]*Institute for nuclear research RAS,*
*117312 Moscow, Russia*
[c]*Imperial College,*
*London, United Kingdom, on leave from ITEP, Moscow, Russia*
*E-mail:* Serguei.Bitioukov@cern.ch

ABSTRACT: We propose an approach for testing the hypothesis that two realizations of the random variables in the form of histograms are taken from the same statistical population (i.e. that two histograms are drawn from the same distribution). The approach is based on the notion "significance of deviation". Our approach allows also to estimate the statistical difference between two histograms.

KEYWORDS: Significance of deviation; Histogram; Normal distribution.

*Corresponding author.

## Contents



## 1. Introduction

The problem of the testing the hypothesis that two histograms are drawn from the same distribution is a very important problem in many scientific researches. For example, this problem exists for the monitoring of the experimental equipment in an experiment. Several approaches to formalize and resolve this problem were considered [1]. Recently, the comparison of weighted histograms was developed in paper [2].

In this note we propose a method which allows to estimate the value of statistical difference between histograms.

## 2. Significance of deviations

In paper [3] several types of significances of deviation (or significance of an enhancement [4]) between two values were considered:

A. significance of deviation between two expected realizations of random variables (for example, $S_{c12}$ [3]);

B. significance of deviation between the observed value and expected realization of random variable (for example, $S_{cP}$ [3]);

C. significance of deviation between two observed values.

As shown (in particular, in paper [3]), many of these significances obey the distribution close to the standard normal distribution if both values are taken from the same statistical population. This property is used here for the estimation of statistical difference between two histograms. We consider the significance of type C in this note.

## 3. Model

Let us consider a simple model with two histograms where the random variable in each bin obeys the normal distribution

$$\varphi(x|n_{ik}) = \frac{1}{\sqrt{2\pi}\sigma_{ik}}\, e^{-\frac{(x-n_{ik})^2}{2\sigma_{ik}^2}}. \tag{3.1}$$

Here the expected value in the bin $i$ is equal to $n_{ik}$ and the variance $\sigma^2_{n_{ik}}$ is also equal to $n_{ik}$. $k$ is the histogram number ($k = 1,2$).

We define the significance as

$$\hat{S}_i = \frac{\hat{n}_{i1} - \hat{n}_{i2}}{\sqrt{\hat{\sigma}^2_{n_{i1}} + \hat{\sigma}^2_{n_{i2}}}}. \tag{3.2}$$

Here $\hat{n}_{ik}$ is an observed value in the bin $i$ of the histogram $k$ and $\hat{\sigma}^2_{n_{ik}} = \hat{n}_{ik}$.

This model can be considered as the approximation of Poisson distribution by normal distribution. So, we suppose that the values $\hat{n}_{ik}$, $(i = 1,2,\ldots,M,\ \ k = 1,2)$ are the numbers of events appeared in the bin $i$ for the histogram $k$. We consider the *RMS* (the root mean square) of the distribution of the significances

$$RMS = \sqrt{\frac{\sum_{i=1}^{M}(\hat{S}_i - \bar{S})^2}{M}}.$$

Here $\bar{S}$ is a mean value of $\hat{S}_i$. The *RMS* has the meaning of the "distance measure" between two histograms. Note that the observed value of the *RMS* can be converted to the $p-$value. If total number of events $N_1$ in the histogram 1 and total number of events $N_2$ in the histogram 2 are various then the normalized significance is used

$$\hat{S}_i(K) = \frac{\hat{n}_{i1} - K\hat{n}_{i2}}{\sqrt{\hat{\sigma}^2_{n_{i1}} + K^2\hat{\sigma}^2_{n_{i2}}}}, \tag{3.3}$$

where $K = \dfrac{N_1}{N_2}$.

Let us consider several examples.

## 4. Examples

All calculations, Monte Carlo experiments and histograms presentation in this note are performed using ROOT code [5]. The number of the bins $M$ is equal to 1000. Histograms are obtained from independent samples.

### 4.1 Uniform distribution

Consider the case when expected values $n_{i1}$ in the first histogram is 66 and the expected values $n_{i2}$ in the second histogram is 45 for each bin number $i = 1,2,\ldots,M$. The results of the Monte Carlo experiment for this example are presented in Fig. 1.

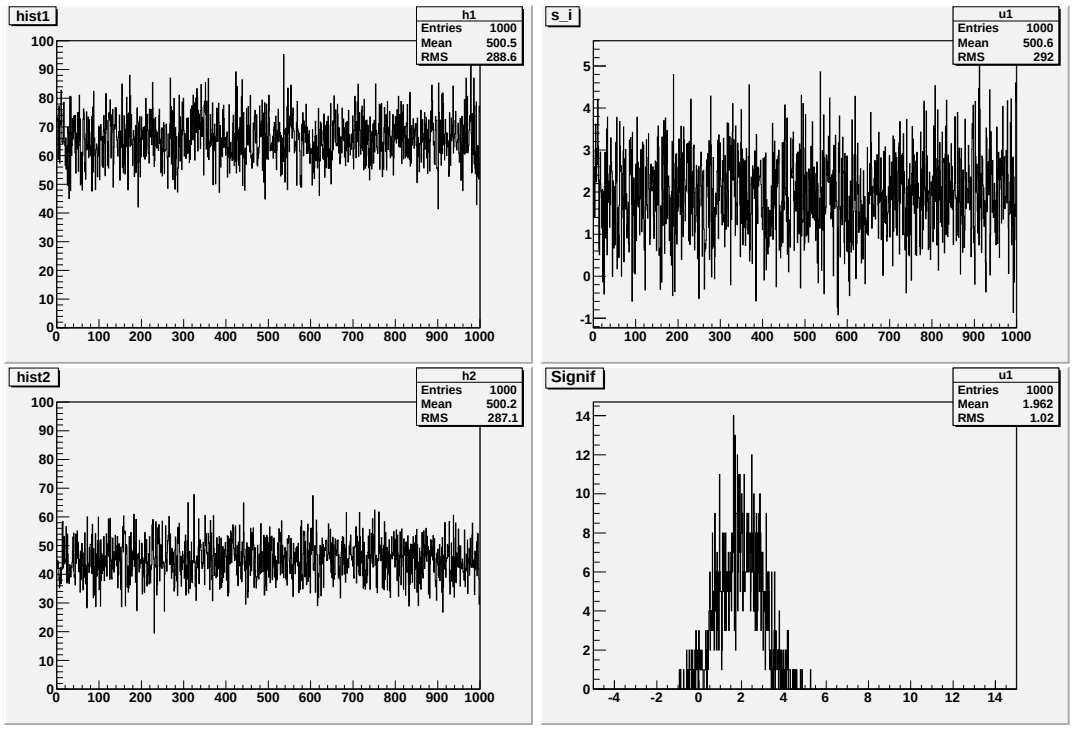


**Figure 1.** Uniform distributions: the observed values $\hat{n}_{i1}$ in the first histogram (left,up), the observed values $n_{i2}$ in the second histogram (left, down), observed significances $S_i$ bin-by-bin (right, up), the distribution of observed significances (right down).

One can see that the distribution of observed significances is close to normal distribution with the $RMS \sim 1$. The average significance is $\sim 1.96$, because total numbers of events in the histograms are different.

In Fig. 2 the corresponding histograms for normalized significances are shown.

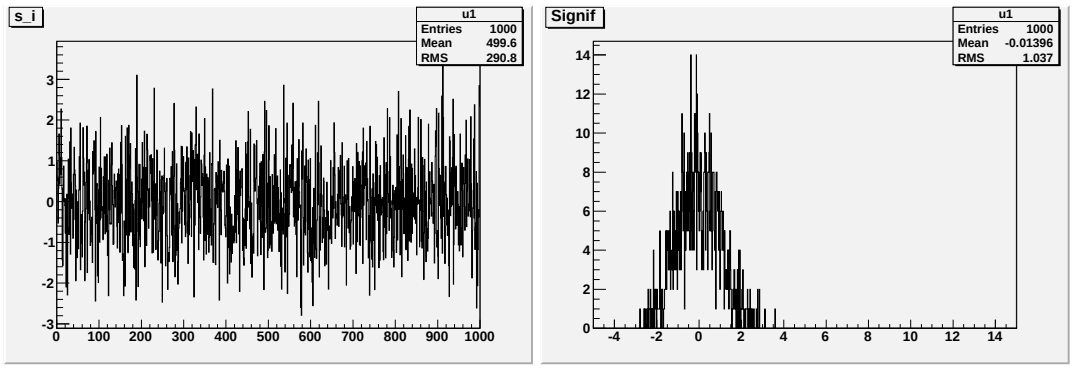


**Figure 2.** Uniform distributions: observed normalized significances $S_i$ bin-by-bin (left), the distribution of observed normalized significances (right).

The distribution of observed normalized significance is close to standard normal distribution.

### 4.2 Triangle distribution

Consider the case when the expected values $n_{ik}$ in both histograms are equal to $i$, where $i$ ($i = 1,2,\ldots,M$) is a bin number and $k$ is a histogram number ($k = 1,2$). It means that the rates of events in different bins are different. One can find that in this case the distributions of observed significances is also close to standard normal distribution, see Fig. 3. It means that the histograms which have different expected values of events in different bins give the distribution of significances close to standard normal distribution.

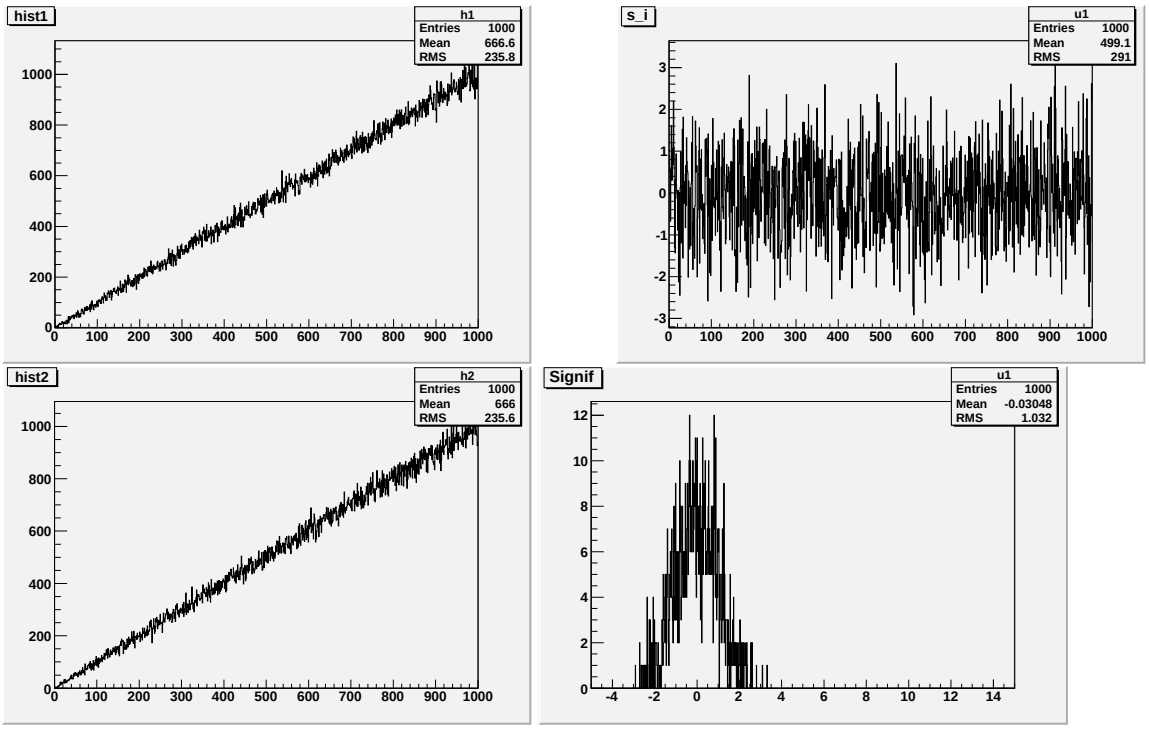


**Figure 3.** Triangle distributions: the observed values $\hat{n}_{i1}$ in the first histogram (left,up), the observed values $n_{i2}$ in the second histogram (left, down), observed significances $S_i$ bin-by-bin (right, up), the distribution of observed significances (right down).

Suppose, the histograms are taken from experiments with different integrated luminosity, i.e. the total numbers of events in histograms are different. In this case the observed significances are changed from bin to bin (see, Fig. 4). Correspondingly, the distribution of significances has non-gaussian shape (in contrast with previous distributon of significances (see, Fig. 3)).

For the normalized significance (Eq. 3.3) we have the standard normal distribution of significances (see, Fig. 5).

So, if two histograms are obtained from the same flow of events then the distribution of the normalized significance obeys to the distribution which is close to the standard normal distribution. The *RMS* of the distribution of significances is a measure of statistical difference between two histograms and, correspondingly, between two flows of events. This "distance measure" between two histogram has a clear interpretation:

- $RMS = 0$ – histograms are identical;

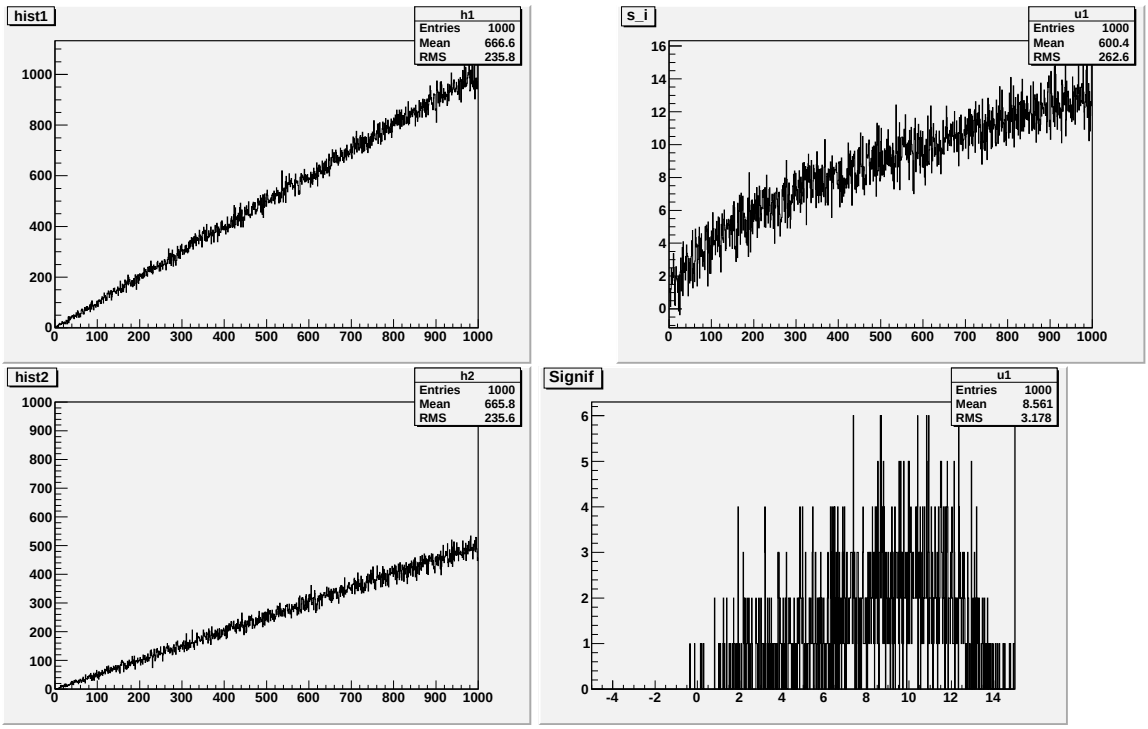


**Figure 4.** Triangle distributions: the observed values $\hat{n}_{i1}$ in the first histogram (left,up), the observed values $n_{i2}$ in the second histogram (left, down), observed significances $S_i$ bin-by-bin (right, up), the distribution of observed significances (right down).

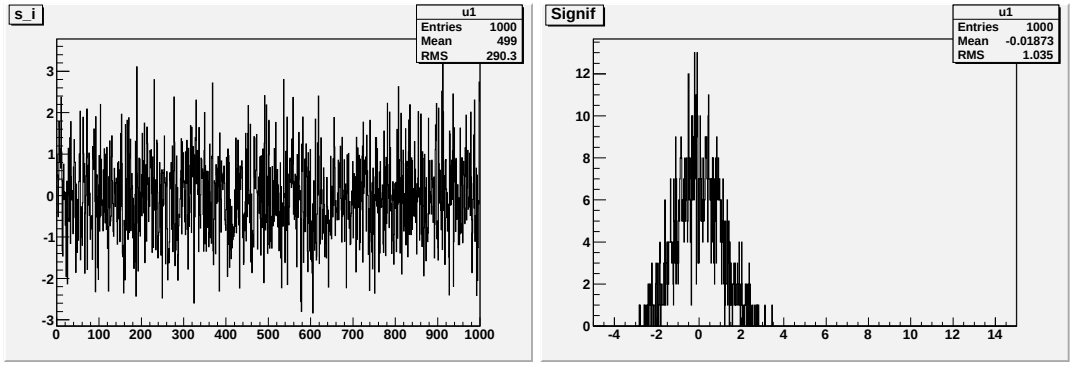


**Figure 5.** Triangle distributions: observed normalized significances $S_i$ bin-by-bin (left), the distribution of observed normalized significances (right).

- $RMS \sim 1$ – both histograms are obtained (by the using independent samples) from the same parent distribution;
- $RMS >> 1$ – histograms are obtained from different parent distributions.

## Acknowledgments

The authors are grateful to L.V. Dudko, V.A. Kachanov, V.A. Matveev and L. Moneta for the

interest and useful comments. The authors would like to thank D. Konstantinov, A. Popov and N. Tsirova for fruitful discussions.

## References

[1] F. Porter, *Testing Consistency of Two Histograms*, arXiv:0804.0380.

[2] N.D. Gagunashvili, *Chi-square tests for comparing weighted histograms*, *Nucl.Instr.&Meth.*, **A614** (2010) 287-296; arXiv:0905.4221.

[3] S.I. Bityukov, N.V. Krasnikov, A.N. Nikitenko, V.V. Smirnova. *Two approaches to Combining Significances*. *Proceedings of Science*, PoS (**ACAT08**) 118.

[4] A.G. Frodesen, O. Skjeggestad, H. Toft, *Probability and Statistics in Particle Physics*, UNIVERSITETSFORLAGET, Bergen-Oslo-Troms, 1979.

[5] R. Brun, F. Rademaker, *ROOT – An object oriented data analysis framework, Nucl.Instr.&Meth.*, **A389** (1997) 81-86.